基于堆稀疏自编码的二叉树集成入侵检测方法 *

柳 毅 [1]，阴梓然 [1]，洪 洲 [2]

(1. 广东工业大学 计算机学院，广州 510006; 2. 广州城市职业学院 科研处，广州 510405)

摘 要：至今已经有许多不同的机器学习方法被提出来，而传统的机器学习方法无法有效解决大规模入侵数据的分类问题，为了解决大规模入侵数据的分类问题，提出的堆稀疏自编码的 lightGBM（light gridient boosting model）二叉树算法。首先将类别标签分为五类，构造成二叉树结构，然后通过上采样方法解决数据分布的不平衡问题，以上处理可以将大规模的数据分解开来以便之后分开训练，再采用稀疏自编码器网络进行特征降维，采用该种降维方法可以保证在原始数据中抽取出更深层特征的基础上节省降维时间。最后通过 lightGBM 集成算法进行分类，而采用 lightGBM 模型相比其他模型可以在保证分类性能的情况下节省训练时间。实验利用 NSL-KDD 数据集测量了所提方法的准确率（accuracy）、精确率（precision）、召回率（recall）以及综合评价指标 F_1 在五类分类上平均分别达到了 87.42%，98.20%，91.31%，优于对比算法，且明显节省了运算时间。

关键词：入侵检测；堆稀疏自编码网络；lightGBM 算法；不平衡数据；NSL-KDD 数据集

中图分类号：TP393.08 doi: 10.19734/j.issn.1001-3695.2018.11.0827

Binary tree ensemble intrusion detection method based on stacked sparse autoencoder

Liu Yi[1], Yin Ziran[1], Hong Zhou[2]

(1. *School of Computer Science & Technology*, *Guangdong University of Technology*, *Guangzhou* 510006, *China*; 2. *Office of Academic Research*, *Guangzhou City Polytechnic*, *Guangzhou* 510405, *China*)

Abstract: So far, many different machine learning methods have been proposed, and traditional machine learning methods can not effectively solve the classification problem of large-scale intrusion data. In order to solve the problem of classification of large-scale intrusion data, This paper proposed lightGBM binary tree algorithm based on stacked sparse autoencoder. Firstly, the category labels were divided into five categories and constructed into binary tree structures, then the imbalance of data distribution was solved by the upper sampling method, the above processing could separate the large-scale data, so that they could be trained separately, and then, the sparse autoencoder network was used to reduce the feature dimension. Using this method could ensure that time of dimension reduction could be saved on the basis of extracting deeper features from the original data. Finally, the lightGBM ensemble algorithm was used to classify. And compared to other models, using the lightGBM model could save training time while ensuring classification performance. The NSL-KDD dataset was used to measure the accuracy, accuracy, recall, and comprehensive evaluation index F1 of the proposed method, which reached an average of 87.42 %, 98.20 %, and 91.31 % in five classification, respectively. It is superior to the comparison algorithm and obviously saves the calculation time.

Key words: intrusion detection; stacked sparse AutoEncoder network; lightgbm algorithm; imbalanced data; nsl-kdd dataset

0 引言

入侵检测是信息安全的重要组成部分，只有正确检测入侵，才能实现后续反应和恢复。入侵检测分为误用检测和异常检测,误用检测通过建模并利用攻击的鲜明特征来检测入侵.误用检测对于已知的入侵具有较高的检测率，但无法检测到新的入侵。异常检测是基于正常行为的模型，任何偏离所构造的正常行为模型的行为都被认为是异常[1]。由于很难对所有正常行为进行精确建模，异常检测很容易将正常行为错误地归类为攻击。

随着移动互联网和物联网的不断发展，网络攻击日益智能化和复杂化，使得恶意入侵的检测更加困难。为了应对这些挑战，机器学习方法在入侵检测中得到了广泛的应用，包括决策树、朴素贝叶斯、随机森林、K-均值聚类算法、支持向量机。传统的浅层结构的机器学习方法大多表达复杂函数的能力有限，泛化能力较弱，因此不能很好地处理复杂的分类问题。

近年来，深度学习已成为机器学习中的一个热门话题，深度学习方法在人脸识别、语音识别、图像识别等领域得到了广泛的应用。同时，在入侵检测中也采用了深度学习方法。文献[2]提出了基于 PCA 降维 KNN 作为分类器的入侵检测方法，实验表明，在应对多分类问题时，对于少数类的检测率明显不高，由此可以看出在应对大规模数据时，对于少数类的处理尤为重要。文献[3]提出了一个半监督约束玻尔兹曼机器（DRBM）模型，它可以检测未知的入侵事件，他们在网络异常检测中的准确率达到了 96%，但应对大规模数据时，

收稿日期：2018-11-15；**修回日期：**2018-12-13 **基金项目：**国家自然科学基金资助项目（61572144）；广州市教育系统创新学术团队资助项目（1201610027）

作者简介：柳毅（1976-），男，江苏连云港人，教授，博士，主要研究方向为网络与信息安全（yiliu@gdut.edu.cn）；阴梓然（1995-），男，湖南岳阳人，硕士研究生，主要研究方向为信息安全；洪洲（1979-），男，江西东乡人，教授，博士，主要研究方向为物联网机器人.

噪声对其模型的影响较大，因此此方法缺乏去噪策略。文献[4]提出了 DBN 和 SVM 的组合，DBN 用于降低输入数据集的维度，SVM 用于分类。这一组合取得了良好的效果，然而，他们没有考虑类别不平衡问题，这对于大规模且种类繁多的入侵数据的侦测性能是致命的。文献[5]表明，堆去噪自编码网络能够很好地区分恶意和非恶意软件，作者构建了三个隐藏的深层神经网络，这个模型只使用了在 SDN 环境中容易获得的六个基本特征，对于特征的考虑过少，容易丢失信息，不适用于入侵检测在数据规模大而繁杂的情况下使用。文献[6]使用一个隐藏层的 RBM 来进行无监督的特征降维。权重被传递给另一个 RBM 产生一个 DBN。预先训练的权重被传递到一个精细的调节层，由一个逻辑回归分类器（用 10 个迭代来训练）与 softmax 层组成。使用 KDD CUP99 组数据对所提出的解决方案进行了评估。作者声称检测率 97.90%，假阴性率为 2.47%。这比类似论文作者声称的结果有所改进，其权重优化使用的随机梯度下降法，在优化时间上会随着数据的规模增大而增加，因此权重优化算法需进一步更新以应对较大的数据规模。文献[7]提出了一种基于深度学习的方法来建立一个有效而灵活的 NIDS。他们的方法被称为自学（STL），它结合了稀疏的自动编码器和 softmax 回归。他们已经实现了他们的解决方案，并根据基准 NSL-KDD 数据集对其进行了评估。在 2 类和 5 类分类中，分类精度都有一定的提高。然而同样在其权重优化算法上可进一步改善。文献[8]提出堆稀疏自编码网络(SSAE)和 XGBoost 集成算法的组合，SSAE 用来降维输入数据，XGBoost 集成算法用来分类。实验结果表示他们在 5 类分类中的平均 F1 值达到了 91.97%，然而他们采用的 5 层稀疏自编码网络的降维效果和在训练时间上可以进一步改善以应对大规模入侵数据。基于以上工作存在的一些欠缺，有必要提出进一步的方法来解决入侵检测在面对大规模数据时的问题，旨在进一步优化入侵检测算法的运算时间，和弥补类别数据不平衡的缺陷，从而改善分类性能。

本文使用了 NSL-KDD 数据集[9]，并采用 Adam 函数作为优化器的稀疏自编码网络进行降维，通过大量的实验，选择了适当数量的前训练迭代数和隐藏层。最后通过 lightGBM 算法(lgb)[10]构造的二叉树结构方法进行分类。贡献在于：采用 Adam 函数作为训练优化器的 3 层堆稀疏自编码来进行降维，不仅提高了少数类的分类精度，而且减少了网络训练的时间，与此同时采用 lightGBM 算法作为分类器进一步减少了训练时间和提高了分类效果。之后采用二叉树的结构并通过 1∶1 比例的上采样方法解决了大部分算法对于少数类侦测率低的缺陷。

1 堆稀疏自编码网络

1.1 单层稀疏自编码（SAE）

如图 1 所示，单层的 SAE 具有输入层、隐藏层和输出层，并且能够给出输入向量的压缩表示，因为隐藏节点的数目小于输入向量的长度。训练过程保证输出向量足够接近输入向量 $X=\{x_1,x_2,...,x_n\}, x_i\in R^n$，$R^n$ 表示输入向量的集合，如此隐藏节点能够表征数据集的有效特征表示。给定一个输入向量 X，所有隐藏节点 $H_j, j=1,2,...s$ 的激励函数计算如下：

$$a(x)=f(W(1)x+b_x) \tag{1}$$

其中：$f(z)=1/(1+\exp(-z))$ 是 sigmoid 激励函数，$W(1)$是关联输入层和隐藏层的权重矩阵，b_x 是输入层的偏置向量，a(x)是包含有 s 个隐藏节点的隐层的激励函数。输出向量 $\hat{x}$，其计算如下：

$$\hat{x}=f(W(1)^T a(x)+b_h) \tag{2}$$

其中：$W(1)^T$ 是关联隐藏层和输出层的权重矩阵，$b_h(1)$是隐藏层的偏置向量。成本函数 J_{SAE} 包括所有输入数据和输出数据之间的误差，权重衰减项和稀疏惩罚项。具体来说，成本函数的定义如下：

$$J_{SAE}=J_{AE}+\beta\sum_{j=1}^{n_h}KL(\rho\|\tilde{\rho}_j) \tag{3}$$

其中：J_{AE} 是不考虑稀疏性时的成本函数，第二项是稀疏惩罚项。具体地说，β 是控制稀疏惩罚项的权重的系数，$KL(\rho\|\tilde{\rho}_j)$ 是 $\tilde{\rho}_j$（隐藏节点 H_j 相对于所有输入数据的平均激励值），和 ρ（设定的稀疏参数）之间的 Kullback-Leibler 散度，其中 n_k 为编号为 k 的隐藏层的节点数，KL 散度其计算为 $\rho\log(\rho/\tilde{\rho}_j)+(1-\rho)\log((1-\rho)/(1-\tilde{\rho}_j))$。值得注意的是隐藏节点的大多数激励结果都限制在接近 0 的值上。第一项 J_{AE} 计算如下：

$$J_{AE}=\frac{1}{2n}\sum_{i=1}^{n}\|\hat{x}_i-x_i\|^2+\frac{\lambda}{2}\sum_{l=1}^{n_l}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}(W_{ij}(l))^2 \tag{4}$$

其中，第一项测量输入和输出数据之间的总误差，其中 n 是输入数据的维数；第二项是权重衰减项，它控制权重的数量，以防止自编码器过拟合。其中 λ 是规范化参数，n_l 是层数，l 是当前层的编号，s_l 和 s_{l+1} 是相邻两个层之间各自隐藏层的节点数量，$W_{ij}(l)$ 是关联相邻两层之间的权重矩阵。

此处优化方法主要通过利用 Adam 算法[11]优化成本函数 J_{SAE} 的权重和偏置值来完成 SAE 的训练过程。Adam 算法的基本原理是将 Momentum 和 RMSprop 结合在一起。经过一定次数的迭代后，J_{SAE} 会被减少到相当小的值，因此实现了 SAE 的自动特征提取。

使用 Adam 算法更新 W 和各层偏置值 b，其具体步骤是首先计算 momentum 指数加权平均数，公式如下：

$$V_{\partial w}=\beta_1 V_{\partial w}+(1-\beta_1)\partial w \tag{5}$$

$$V_{\partial b}=\beta_1 V_{\partial b}+(1-\beta_1)\partial b \tag{6}$$

使用 RMSprop 算法进行更新，公式如下：

$$S_{\partial w}=\beta_2 S_{\partial w}+(1-\beta_2)(\partial w)^2 \tag{7}$$

$$S_{\partial b}=\beta_2 S_{\partial b}+(1-\beta_2)(\partial b)^2 \tag{8}$$

此时将式（5）～（8）皆考虑修正偏差，公式如下：

$$V_{\partial w}^{c}=\frac{V_{\partial w}}{(1-\beta_1^t)} \tag{9}$$

$$S_{\partial w}^{c}=\frac{S_{\partial w}}{(1-\beta_2^t)} \tag{10}$$

$$V_{\partial b}^{c}=\frac{V_{\partial b}}{(1-\beta_1^t)} \tag{11}$$

$$S_{\partial b}^{c}=\frac{S_{\partial b}}{(1-\beta_2^t)} \tag{12}$$

最终参数的更新函数如下：

$$W_{ij}(l)=W_{ij}(l)-\alpha\frac{V_{\partial w}^{c}}{\sqrt{S_{\partial w}^{c}+\varepsilon}} \tag{13}$$

$$b_i(l)=b_i(l)-\alpha\frac{V_{\partial b}^{c}}{\sqrt{S_{\partial b}^{c}+\varepsilon}} \tag{14}$$

其中：α 为学习率，V 表示移动均值，S 表示平方梯度，β_1、β_2 是指数衰减率，ε 是设定的步长，t 表示某个时刻，而 β_1^t 和 β_2^t 表示在 t 时刻的相应值。

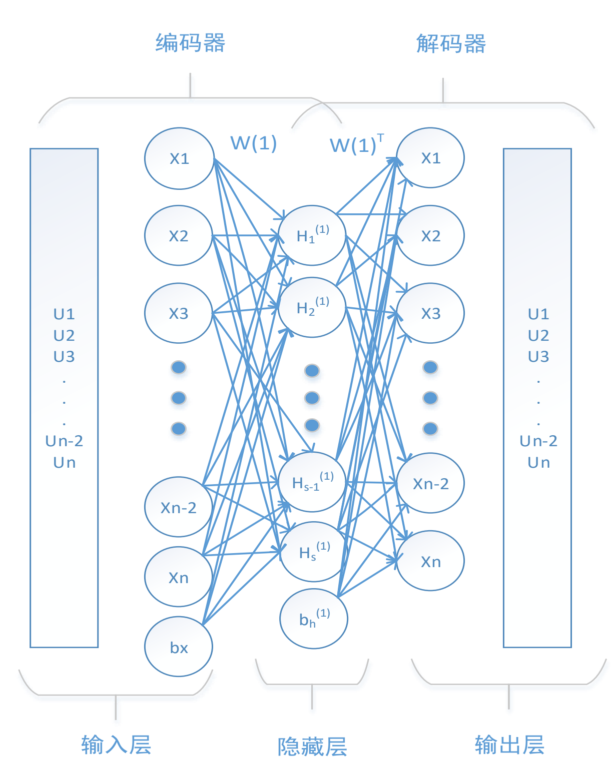


图 1　稀疏自编码结构

Fig. 1　Architecture of SAE

1.2　堆稀疏自编码（SSAE）

顾名思义，SSAE 是一种分层编码结构，单层 SAE 被堆积起来。每个隐藏层都希望从上一层学习到更抽象的特征表示[12]。SSAE 中每层的训练过程与 SAE 的训练过程是相同的，即最大限度地降低成本函数，并在每次获得一层最优权值和偏置值。在所有的层经过适当的训练后，SSAE 能够从训练数据集的输入数据中学习到更复杂和抽象的特征表示。

图 2 是堆稀疏自编码的训练结构图，因堆稀疏自编码的构建方式为先单独训练各个 SAE，每一层的输入皆是其上一层的隐层，最后有序的将各级隐层连接起来构建成整体的 SSAE。第一层由 x、h_2、X 组成再使用式(4)来无监督的学习特征表示，之后使用式(13)(14)得到权重和偏置 W1、b1，第二层由 h_1，h_2，H_2 组成，由训练过层与第一层相同并得到 W2、b2，重复以上步骤最终得到整个网络的参数。

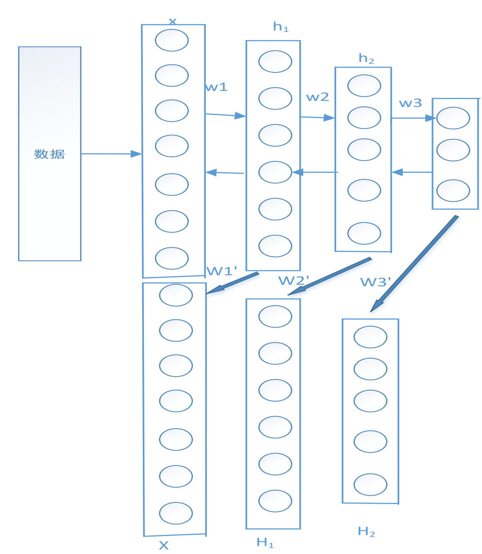


图 2　堆稀疏自编码的训练结构图

Fig. 2　Training architecture of SSAE

2　基于堆稀疏自编码的二叉树集成算法

2.1　数据处理流程

如图 3 所示，首先导入 NSL-KDD 数据集；其次，将数据作为神经网络的输入，因此需要对数据进行预处理。对于连续特征，需要规范化来平衡每一维特征的影响，而对于类别特征则需要独热编码。然后，标准化数据被用作 SSAE-lightGBM 二叉树的输入。最后，利用该模型对实验数据进行了预测，并对实验结果进行了对比分析。

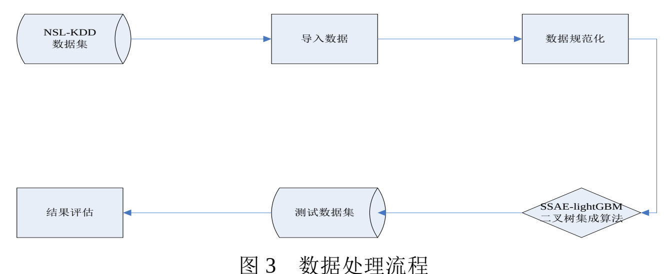


图 3　数据处理流程

Fig. 3　Process of data processing

2.2　堆稀疏自编码的 lightGBM 二叉树结构算法(SSAE-lgb-BT)框架

在入侵检测中，数据类别的不平衡问题频繁发生，造成了对数据量少的类别预测性能不佳的效果。解决类别数据量不平衡问题的方法一般包括上采样、下采样、代价敏感学习算法。通过参考决策树的分类过程和对入侵数据集分布的分析，引入了二叉树来解决入侵检测问题。通过使用二叉树，将多重分类简化为二元分类，使原来的多分类中类别数据的失衡问题变得相对平衡，减少了接下来集成方法计算的次数。

二分类可能还存在数据不平衡问题。在数据层面，过采样和欠采样是最具代表性的方法。过采样法为数据量少的类别创造合成的样本，并将其添加到训练集中，这将需要相当长的时间来进行训练。通过减少数据量多的类别样本的数量，欠采样方法通过平衡不同类别的比例，但这可能会丢失一些重要的信息。基于以上分析，提出了一种将 EUS[13] 和 SMOTE[14]混合的方法 EUS-SMOTE。EUS-SMOTE 方法先使用 EUS 方法将多数类与存在多个类别的少数类的样本集分

开，然后使用 SMOTE 方法对少数类进行采样，最后将多数类样本与过采样后的样本结合作为训练集，之后再重复将存在多个类别的少数类采用相同的 EUS-SMOTE 方法。关于不平衡类别数据的过采样比例使用 1：1 进行数据的过采样。最后在每层分开的数据集上使用 SSAE- lightGBM 集成算法进行训练并进行分类，其算法结构框架如图 4 所示。

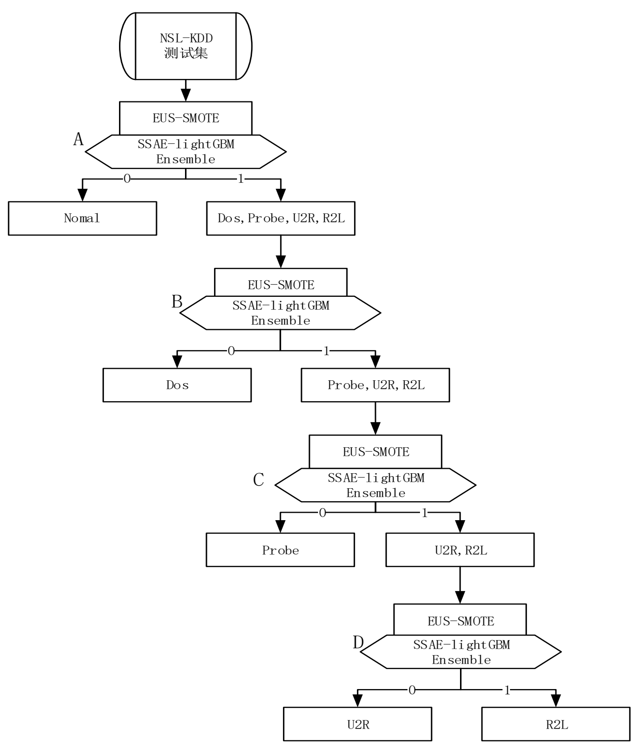


图 4　算法框架

Fig. 4　Structure of algorithm

SSAE-lightGBM 集成算法的训练过程如图 5 所示，图中(a)~(d)训练的最终分类器依次对应图 4 中 A、B、C、D 四个分类器。图 5(a)中，首先将 NSL-KDD 训练集分为两大类，分别为 Nomal 类以及(Dos，Probe，U2R，R2L)类，其标签分别标记为 0、1，之后使用 EUS-SMOTE 处理 0、1 两类的类别平衡问题，之后利用 SSAE 抽取数据的深层特征 T，最后将 T 放入 lightGBM 模型训练，训练过程使用 bagging 投票式的 5 折交叉验证法融合预测结果，并最终得到分类器 A。同样通过以上相同步骤得到分类器 B、C、D。

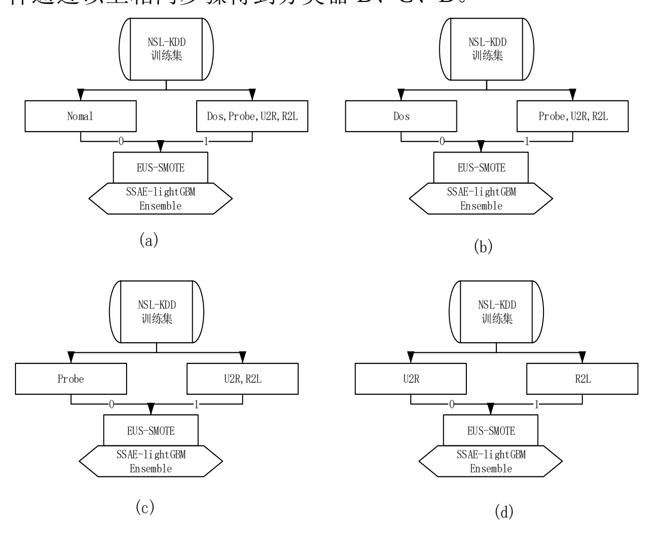


图 5　各分类器的训练流程

Fig. 5　Training process of various classifier

图 4 所示为预测阶段，入侵数据首先进入第一层分类器 A，通过 SSAE-lightGBM 集成模型后将会被分为 0 类或 1 类，若被判为 0 类则输出分类结果，并转换为真实的标签，若被分为 1 类则会将原始的侵入数据输入到第二层分类器 B 处分类，之后重复以上操作直到结束。

图 6 是使用 SSAE-lightGBM 模型预测入侵数据时的详细过程。其前半段是由 3 层训练好的隐层组成的编码器（训练过程如图 2 所示），当数据依次通过一~三层时，原始特征维数将会由下一层的神经元数所确定，直至到达最后一层，编码器将从高维特征中自动抽取出有意义的特征。之后将特征作为 lightGBM 分类器的输入从而得到预测结果。关于图 6 前半部分的编码器一共有 4 个，分别是在图 5 训练过程中产生，此 4 个编码器分别是从不同是数据集中无监督学习产生。

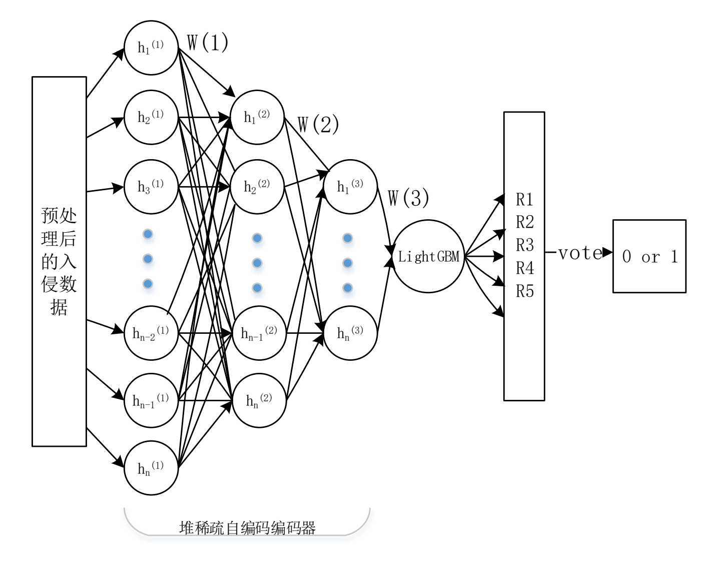


图 6　模型预测过程

Fig. 6　Predicting process of model

3　实验与结果

为了证明所提算法的有效性，本文使用 NSL-KDD 公用数据集将所提出的 SSAE-lightGBM 二叉树结构算法与 SSAE-XGBoost 二叉树结构算法、PCA-XGBoost-二叉树结构算法的分类效果进行比较。以上几种方法在 Intel(R) Core™ i5-3210M CPU @ 2.5Hz 处理器、4 GB 内存、Windows7 64 位操作系统和 Pycharm2017 的环境中运行。

3.1　数据及描述

对所有的 NSL-KDD 数据集进行了实验。NSL-KDD 数据集由 KDDcup99 数据集生成。它解决了 KDDcup99 数据集中的数据冗余问题，更具实用性。NSL-KDD 数据集包含 125 973 训练样本和 22 544 测试样本，它包括四类攻击：拒绝服务攻击（DoS），远程到本地攻击（R2L），用户根目录攻击（U2R）和攻击者试图获取有关目标主机信息的嗅探攻击（Probe）。图 7 显示了训练数据和测试数据的分布情况，数据在不同类别中的分布是不平衡的。DoS 类别的数目比 U2R 类别的数目多得多。

NSL-KDD 数据集包含 41 个特性和 1 个类标签。这 41 个特征包含 38 个连续特征和 3 个类别特征。

3.2　评估方法

采用精确度、召回率和 F1 值计量方法。用这些方法来比较不同模型的结果。评估措施的计算公式如下：

$$Recall=\frac{TP}{TP+FN}\times 100\% \tag{15}$$

$$Precision = \frac{TP}{TP+FP} \times 100\% \quad (16)$$

$$F1 = \frac{2TP}{2TP+FP+FN} \times 100\% \quad (17)$$

其中：TP（真阳率）表示正常类型数据被正确归类的情况，TN（真阴率）表示攻击类型数据被正确归为攻击类别的情况，FN（假阴性）表示将正常类型归为攻击类型的情况，而 FP（假阳性）指的是攻击类型被归类为正常类型的情况。精确度指出返回的正常类别有多少是正确的，而召回率则指出模型返回的攻击有多少是被分错的。F_1 测度是精确度和召回率的调和平均值。

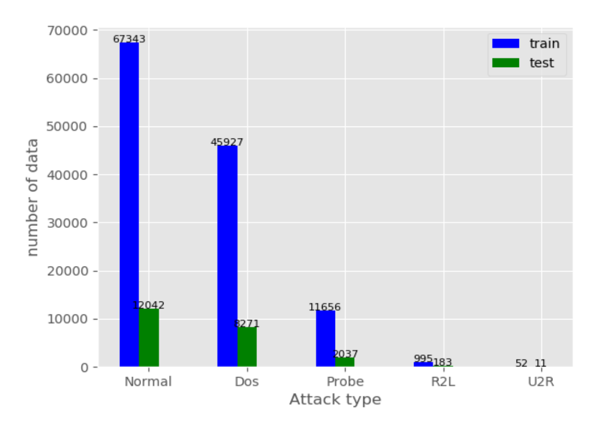


图 7 训练数据集与测试数据集的分布

Fig. 7 Distribution of data of train and test

3.3 稀疏自编码的迭代数及隐藏层数的选择

对于 SSAE-lightGBM 的构成要素来说，预训练迭代的数量和隐藏层的数量非常重要。预训练迭代次数太少则无法减少损失，而太多则浪费机器资源。过多的隐藏层会导致过度拟合，过少的隐藏层无法达到良好的检测性能。因此，对预训过程中的迭代次数和隐藏层数进行了实验，该实验选取的 batch_size 大小为 64，epochs 为 80，隐藏层选取 100-80-60 的三层结构。

在不同隐藏层的情况下，损失与预训练迭代次数的关系如图 8 所示。从图 8 可以看出，随着迭代的增加,不同隐藏层的损失呈不同的下降趋势。随着迭代的增加，第 1、2 和第 3 层的损失下降在 40 迭代次数后都趋于稳定。通过以上分析本文选择了含 3 个隐藏层和 40 个预训练迭代数的网络结构。

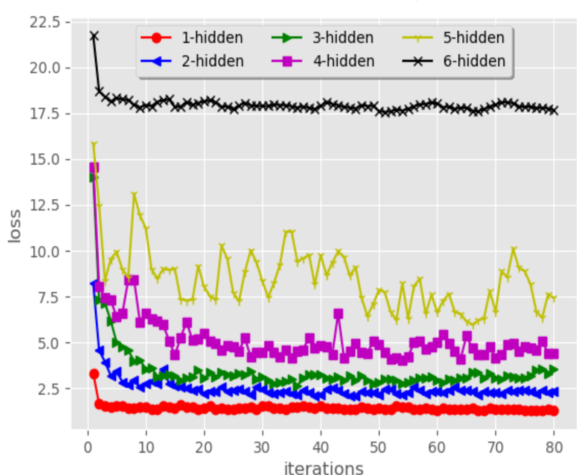


图 8 预训练迭代数、隐藏层数与损失的关系

Fig. 8 Relationship between pre-training, hidden number and loss

3.4 实验性能评估

表 1 展示了堆稀疏自编码的 lightGBM 二叉树结构算法（SSAE-lgb-BT ensemble）、堆稀疏自编码的极限梯度提升集成二叉树结构算法(SSAE-xgb-BT ensemble)、主成分分析的 lightGBM 二叉树结构算法（PCA-lgb-BT）分别在 NSL-KDD 数据集上的性能表现,将三种算法分别依次用 SLB、SXB、PLB 表示。F_1 值反映了模型的检测效果。在 F_1 值方面，表中显示 SSAE-xgb-BT 算法优于不进行深度特征提取 PCA-xgb-BT 算法,并且从运行时间上，也只花费了后者 1/4 的时间，从这可以看出堆稀疏自编码在提取特征方面要优于主成分分析法。从 SSAE-lgb-BT 和 SSAE- xgb-BT 算法的性能可以看出，在 Probe 类上 F_1 值达到了 97.09%，高出后者 9 个百分点。虽然在少数类 R2L 上精确度偏低，但与 SSAE-xgb-BT 算法相比，对于少数类 U2R，在召回率上有明显提升，并且在计算时间上，SSAE-lgb-BT 集成算法只用了 SSAE-xgb-BT 集成算法的将近 1/5 的时间，从而在整体性能表现上，本文提出的混合分类器的性能优于 SSAE-xgb-BT 等其他分类算法。

表 2 展示了 DNN(1)[7]、DNN(2)[15] 两种算法与 SSAE-lgb-BT 集成算法的比较，上述两种算法的数据都出自于论文中所给实验结果。从结果比对可以看出，SSAE-lgb-BT 集成算法作为一种混合算法要优于 DNN 算法的，由此可以进一步说明使用堆稀疏自编码进行特征降维可以从高维特征中学习到更深层次的特征。根据 F1 值，与所提算法进行比较，可以看出 SSAE-lgb-BT 算法可以更好的处理数据不平衡问题，从而有更好的分类效果。

表 1 不同模型的结果比较

Table 1 Comparison of different model

模型	类别	精确度(%)	召回率(%)	F_1(%)	time(s)
SLB	Normal	98.64	98.82	98.73	673
	Dos	99.01	99.15	99.08	
	Probe	97.54	96.64	97.09	
	R2L	73.77	96.42	83.59	
	U2R	68.18	100.00	78.09	
SXB	Normal	97.96	99.96	98.94	3232
	Dos	98.57	99.17	98.86	
	Probe	75.75	99.50	86.02	
	R2L	99.18	97.13	98.14	
	U2R	69.26	89.00	77.89	
PLB	Normal	90.31	92.99	91.63	14987
	Dos	98.75	22.85	37.11	
	Probe	20.07	80.51	32.13	
	R2L	13.73	3.83	2.99	
	U2R	25.00	18.18	21.05	

表 2 与其他模型比较

Table 2 Comparison with other models

模型	精确度(%)	召回率(%)	F_1(%)
SSAE-lgb-BT ensemble	87.42	98.20	91.31
DNN(1)	83.00	69.00	75.35
DNN(2)	83.00	75.00	74.00

4 结束语

利用堆稀疏自编码网络，以无监督的方式学习入侵检测数据的深层特征。稀疏性约束增强了堆稀疏自编码网络的泛化能力。实验结果表明，本文提出的 SSAE-lgb 方法能够从高维入侵数据中提取出深层的稀疏特征。与线性降维法主要成分分析法（PCA）相比，SSAE-lgb-BT 集成算法显著提高了检测效果，与运用随机梯度提升法作为优化器的 5 层堆稀

疏自编码网络 SSAE-xgb-BT 集成算法相比，SSAE-lgb-BT 集成算法进一步提升了准确率，并节约了更多的计算时间,虽然在少数类 R2L 上精确度偏低，但与 SSAE-xgb-BT 算法相比，对于少数类 U2R，在召回率上有明显提升，从而在整体性能表现上。所以本文提出的混合分类器的性能优于其他分类算法，F_1 值达到了平均 91.31%。并且，本文的方法能够很好地处理类别失衡问题，提高少数类别的 F_1 值。因此本文为网络入侵检测提供了一种新的研究方法。在接下来的工作中将进一步提升算法的性能，本文通过将训练集数据进行聚类（聚类中心为 5），再将测试集对 5 个聚类中心进行分类（可采用 KNN），如此可以进一步提高 SSAE-lgb 二叉树集成算法的准确率和计算时间。

参考文献：

[1] Sarasamma S T, Zhu Qiuming A, Huff J. Hierarchical Kohonenen net for anomaly detection in network security [J]. IEEE Trans on Systems Man & Cybernetics Part B Cybernetics, 2005, 35(2): 302.

[2] Khalid C, Zyad E, Mohammed B. Network intrusion detection system using L1-norm PCA [C]//Proc of the 11th International Conference on Information Assurance & Security. 2016.

[3] Fiore U, Palmieri F, Castiglione A, *et al*. Network anomaly detection with the restricted Boltzmann machine. [J]. Neurocomputing, 2013, 122: 13-23.

[4] Salama M A, Eid H F, Ramadan R A, *et al.* Hybrid Intelligent Intrusion Detection Scheme [M]//Soft Computing in Industrial Applications, Volume 96 of the Advances in Intelligent and Soft Computing Book Series. Berlin: Springer-Verlag, 2011: 293-303.

[5] Wang Yao, Cai Wandong, Wei Pengcheng. A deep learning approach for detecting malicious JavaScript code [J]. Security & Communication Networks, 2016, 9(11): 1520-1534.

[6] Alrawashdeh K, Purdy C. Toward an online anomaly intrusion detection system based on deep learning [C]//Proc of the15th IEEE International Conference on Machine Learning & Applications. Piscataway,NJ: IEEE Press,2017.

[7] Javaid A Y, Niyaz Q, Sun Weqing, *et al.* A deep learning approach for network intrusion detection system [C]// Proc of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies. New York: ACM Press, 2015: 21-26.

[8] Zhang Baoan, Yu Yanhua, Li Jie. Network intrusion detection based on stacked sparse autoencoder and binary tree ensemble method [C]//Proc of IEEE International Conference on Communications Workshops. Piscataway,NJ: IEEE Press, 2018: 20-2.

[9] Tavallaee M, Bagheri E, Lu Wei, *et al.* A detailed analysis of the KDD CUP 99 data set [C]// Proc of the 2nd IEEE International Conference on Computational Intelligence for Security & Defense Applications. Piscataway, NJ: IEEE Press, 2009: 53-58.

[10] Ke Guolin, Meng Qi, Thomas F, LightGBM: a highly efficient gradient boosting decision tree [C]// Proc of the 31st Conference on Neural Information Processing Systems. Long Beach, CA: NIPS Press, 2017.

[11] Kingma D P, Ba J. Adam: a method for stochastic optimization [C]// Proc of the 3rd International Conference for Learning Representations. 2015.

[12] Lee H, Grosse R, Ranganath R, *et al.* Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations [C]// Proc of the 26th Annual International Conference on Machine Learning. New York: ACM Press, 2009: 609–616.

[13] Chawla N V, Bowyer K W, Hall L O, *et al.* SMOTE: synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321-357.

[14] Garcia S, Herrera F. Evolutionary undersampling for classification with imbalanced datasets: proposals and taxonomy [J]. Evolutionary Computation, 2014, 17(3): 275-306.

[15] Tang T A, Mhamdi L, McLernon D, *et al.* Deep learning approach for network intrusion detection in software defined networking [C]//Proc of International Conference on Wireless Networks & Mobile Communications. Piscataway,NJ: IEEE Press, 2016.